Ridge Regression: Regularizing Linear Models

Nipun Batra

IIT Gandhinagar

July 30, 2025

Outline

When Linear Regression Goes Wrong

Problem: What happens when we have too many features or complex polynomials?

High-Degree Polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_n x^n$$

When Linear Regression Goes Wrong

Problem: What happens when we have too many features or complex polynomials?

High-Degree Polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_n x^n$$

When Linear Regression Goes Wrong

Problem: What happens when we have too many features or complex polynomials?

High-Degree Polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_n x^n$$

Issues:

- Large coefficients $|c_i|$ can cause instability

When Linear Regression Goes Wrong

Problem: What happens when we have too many features or complex polynomials?

High-Degree Polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_n x^n$$

Issues:

- Large coefficients $|c_i|$ can cause instability
- Overfitting to training data

When Linear Regression Goes Wrong

Problem: What happens when we have too many features or complex polynomials?

High-Degree Polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_n x^n$$

Issues:

- Large coefficients $|c_i|$ can cause instability
- Overfitting to training data
- Poor generalization to new data

When Linear Regression Goes Wrong

Problem: What happens when we have too many features or complex polynomials?

High-Degree Polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_n x^n$$

Issues:

- Large coefficients $|c_i|$ can cause instability
- Overfitting to training data
- Poor generalization to new data

When Linear Regression Goes Wrong

Problem: What happens when we have too many features or complex polynomials?

High-Degree Polynomial

$$f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots + c_n x^n$$

Issues:

- Large coefficients $|c_i|$ can cause instability
- Overfitting to training data
- Poor generalization to new data

The Key Insight

Solution: Penalize large coefficients to encourage simpler models!

Pop Quiz: Overfitting Intuition

Quick Quiz 1

A linear model with coefficients [0.1, 0.2, -0.1] vs [10.2, -15.6, 23.4]. Which is likely to generalize better?

a) [10.2, -15.6, 23.4] (larger coefficients)

Answer: b) Smaller coefficients typically indicate a more stable, generalizable model!

Pop Quiz: Overfitting Intuition

Quick Quiz 1

A linear model with coefficients [0.1, 0.2, -0.1] vs [10.2, -15.6, 23.4]. Which is likely to generalize better?

a) [10.2, -15.6, 23.4] (larger coefficients)
b) [0.1, 0.2, -0.1] (smaller coefficients)

Answer: b) Smaller coefficients typically indicate a more stable, generalizable model!

Pop Quiz: Overfitting Intuition

Quick Quiz 1

A linear model with coefficients [0.1, 0.2, -0.1] vs [10.2, -15.6, 23.4]. Which is likely to generalize better?

a) [10.2, -15.6, 23.4] (larger coefficients)
b) [0.1, 0.2, -0.1] (smaller coefficients)
c) Both are equivalent

Answer: b) Smaller coefficients typically indicate a more stable, generalizable model!

Two Ways to Think About Ridge Regression

Two Ways to Think About Ridge Regression

Two Ways to Think About Ridge Regression

Constrained Form

$$\text{Minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$
$$\text{subject to } \|\boldsymbol{\theta}\|^2 \leq S$$

Interpretation: Find best fit while keeping coefficients "small"

Two Ways to Think About Ridge Regression

Constrained Form

$$\text{Minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$
$$\text{subject to } \|\boldsymbol{\theta}\|^2 \leq S$$

Interpretation: Find best fit while keeping coefficients "small"

Two Ways to Think About Ridge Regression

Constrained Form

$$\text{Minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$
$$\text{subject to } \|\boldsymbol{\theta}\|^2 \leq S$$

Interpretation: Find best fit while keeping coefficients "small"

Two Ways to Think About Ridge Regression

Constrained Form

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$

subject to $\|\boldsymbol{\theta}\|^2 \leq S$

Interpretation: Find best fit while keeping coefficients "small"

Penalized Form

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$

Interpretation: Balance fit quality vs coefficient size

Two Ways to Think About Ridge Regression

Constrained Form

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$

subject to $\|\boldsymbol{\theta}\|^2 \leq S$

Interpretation: Find best fit while keeping coefficients "small"

Penalized Form

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$

Interpretation: Balance fit quality vs coefficient size

Two Ways to Think About Ridge Regression

Constrained Form

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$

subject to $\|\boldsymbol{\theta}\|^2 \leq S$

Interpretation: Find best fit while keeping coefficients "small"

Penalized Form

Minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$

Interpretation: Balance fit quality vs coefficient size

Key Insight

These formulations are **equivalent**! Different λ values correspond to different constraint budgets S.

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Components:

- **Data Fit**: How well model fits training data

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Components:

- **Data Fit**: How well model fits training data
- **Penalty**: Discourages large coefficients

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Components:

- **Data Fit**: How well model fits training data
- **Penalty**: Discourages large coefficients
- λ **(lambda)**: Controls the trade-off

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Components:

- **Data Fit**: How well model fits training data
- **Penalty**: Discourages large coefficients
- λ **(lambda)**: Controls the trade-off

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Components:

- **Data Fit**: How well model fits training data
- **Penalty**: Discourages large coefficients
- λ **(lambda)**: Controls the trade-off

λ **behavior:**

- $\lambda = 0$: No regularization (ordinary least squares)

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Components:

- **Data Fit**: How well model fits training data
- **Penalty**: Discourages large coefficients
- λ **(lambda)**: Controls the trade-off

λ **behavior:**

- $\lambda = 0$: No regularization (ordinary least squares)
- λ small: Slight regularization

Understanding the Ridge Penalty

Ridge Regression Objective

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}_{\text{Data Fit}} + \underbrace{\lambda\|\boldsymbol{\theta}\|^2}_{\text{Complexity Penalty}}$$

Components:

- **Data Fit**: How well model fits training data
- **Penalty**: Discourages large coefficients
- λ **(lambda)**: Controls the trade-off

λ **behavior:**

- $\lambda = 0$: No regularization (ordinary least squares)
- λ small: Slight regularization
- λ large: Heavy regularization (coefficients shrink toward 0)

Pop Quiz: Ridge Parameter

Quick Quiz 2

What happens to the coefficients as λ increases in Ridge regression?

a) Coefficients become larger to minimize error

Answer: b) Ridge penalty $\lambda\|\boldsymbol{\theta}\|^2$ forces coefficients to shrink!

Pop Quiz: Ridge Parameter

Quick Quiz 2

What happens to the coefficients as λ increases in Ridge regression?

a) Coefficients become larger to minimize error
b) Coefficients shrink toward zero

Answer: b) Ridge penalty $\lambda\|\boldsymbol{\theta}\|^2$ forces coefficients to shrink!

Pop Quiz: Ridge Parameter

Quick Quiz 2

What happens to the coefficients as λ increases in Ridge regression?

a) Coefficients become larger to minimize error
b) Coefficients shrink toward zero
c) Coefficients remain unchanged

Answer: b) Ridge penalty $\lambda\|\boldsymbol{\theta}\|^2$ forces coefficients to shrink!

Lagrangian Approach

Starting from the constrained formulation:

$$\text{Minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$
$$\text{subject to } \|\boldsymbol{\theta}\|^2 \leq S$$

Lagrangian Approach

Starting from the constrained formulation:

$$\text{Minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$
$$\text{subject to } \|\boldsymbol{\theta}\|^2 \leq S$$

Lagrangian:

$$L(\boldsymbol{\theta}, \mu) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \mu(\|\boldsymbol{\theta}\|^2 - S)$$

where $\mu \geq 0$ is the Lagrange multiplier.

Lagrangian Approach

Starting from the constrained formulation:

$$\text{Minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$
$$\text{subject to } \|\boldsymbol{\theta}\|^2 \leq S$$

Lagrangian:

$$L(\boldsymbol{\theta}, \mu) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \mu(\|\boldsymbol{\theta}\|^2 - S)$$

where $\mu \geq 0$ is the Lagrange multiplier.

Key Insight

Setting $\lambda = \mu$ gives us the penalized form! Different constraint budgets S correspond to different penalty strengths λ.